Supplementary Material A

Memory distortion: Less probable contexts result in more error in prediction

Here we give formal derivations of the two deductions from Section 3.4 about the interactions of memory and probabilistic expectations under lossy-context surprisal. The two deductions are:

1. Less probable contexts lead to less accurate predictions and thus more processing difficulty. (Proposition 1)
2. As noise affects memory representations, comprehenders will regress to their prior expectations without regard for context. (Proposition 2)

We will support these claims using bounding arguments.

In order to reason about the effects of memory in lossy-context surprisal it is useful to employ a concept we call **memory distortion**: the extent to which the predictions of lossy-context surprisal diverge from the predictions of a theory with perfect memory, as a function of the memory encoding function M. Memory distortion is simply the difference between $D_{\text{surprisal}}$ and $D_{\text{lc surprisal}}$ as a function of the memory encoding function M for a word w in a context c:

$$\text{distortion}_M(w,c) = D_{\text{lc surprisal}}(w|c) - D_{\text{surprisal}}(w|c).$$

Memory distortion is thus equal to:

$$\begin{aligned}\text{distortion}_M(w,c) &\equiv \mathop{\mathbb{E}}_{r\sim M(c)}\left[-\log p(w|r)\right] - (-\log p(w|c)) &(15)\\ &= \log p(w|c) - \mathop{\mathbb{E}}_{r\sim M(c)}\left[\log p(w|r)\right]\\ &= \mathop{\mathbb{E}}_{r\sim M(c)}\left[\log\frac{p(w|c)}{p(w|r)}\right]. &(16)\end{aligned}$$

In order to support the first deduction, we will show that memory distortion is upper bounded by the information content of the context (Proposition 1).

Proposition 1. *For all conditional distributions M of memory representations given contexts, and all contexts c and all words w,*

$$\textit{distortion}_M(w,c) \le -\log p(c).$$

Proof. First we show that memory distortion for all w, c, and M is upper bounded by a quantity we call the **irrecoverability** of the context c under the memory model M. Irrecoverability answers the question: on average, given a memory representation r drawn from the distribution $M(c)$, how many bits of additional information would be required to recover c with certainty? To show this, we apply the fact that the distribution over the next word w is conditionally independent from the distribution over memory representations r given the true context c. This is an assumption of the model, as indicated in the Bayesian network in Figure 2(b). Symbolically, we have:

$$W \perp R|C,$$

where W is the random variable ranging over words, C is the random variable ranging over contexts, and R is the random variable ranging over memory representations. Using this independence assumption, we can write:

$$\begin{aligned} \text{distortion}_M(w,c) &= \mathop{\mathbb{E}}_{r\sim M(c)}\left[\log\frac{p(w|c)}{p(w|r)}\right] \qquad (16) \\ &= \mathop{\mathbb{E}}_{r\sim M(c)}\left[\log\frac{p(w|c,r)}{p(w|r)}\right]. \end{aligned}$$

Now we use Bayes' rule to rewrite $p(w|c,r)$ as $\frac{p(c|w,r)p(w|r)}{p(c|r)}$ and cancel out the terms $p(w|r)$:

$$\begin{aligned} \text{distortion}_M(w,c) &= \mathop{\mathbb{E}}_{r\sim M(c)}\left[\log\frac{p(c|w,r)\cancel{p(w|r)}}{\cancel{p(w|r)}p(c|r)}\right] \\ &= \mathop{\mathbb{E}}_{r\sim M(c)}\left[\log\frac{p(c|w,r)}{p(c|r)}\right]. \qquad (17) \end{aligned}$$

Now the numerator quantity $p(c|w,r)$ has maximum value 1, since no probability value may exceed 1. Therefore we have:

$$\begin{aligned} \text{distortion}_M(w,c) &= \mathop{\mathbb{E}}_{r\sim M(c)}\left[\log\frac{p(c|w,r)}{p(c|r)}\right] \qquad (17) \\ &\leq \mathop{\mathbb{E}}_{r\sim M(c)}\left[\log\frac{1}{p(c|r)}\right] \qquad (18) \\ &= \mathop{\mathbb{E}}_{r\sim M(c)}\left[-\log p(c|r)\right]. \qquad (19) \end{aligned}$$

The final quantity in Equation 19 here is the irrecoverability of context c on average given memory representations r. This same quantity was used as a measure of the average information content of a word given its context in Piantadosi, Tily, and Gibson (2011), and as a measure of the codability of color stimuli in Gibson et al. (2017).

Finally, we show that the irrecoverability is itself upper bounded by the information content of the context c. This result means intuitively that the maximum number of bits you might require to recover an object does not exceed the number of bits in the total information content of the object itself. To show this formally, we again use Bayes' rule to rewrite $p(c|r)$ as $\frac{p(r|c)p(c)}{p(r)}$:

$$\mathop{\mathbb{E}}_{r \sim M(c)}\left[\log \frac{1}{p(c|r)}\right] = \mathop{\mathbb{E}}_{r \sim M(c)}\left[\log \frac{p(r)}{p(r|c)p(c)}\right].$$

Now we separate out the term $\log \frac{1}{p(c)}$:

$$\mathop{\mathbb{E}}_{r \sim M(c)}\left[\log \frac{p(r)}{p_M(r|c)p(c)}\right] = \log \frac{1}{p(c)} - \underbrace{\mathop{\mathbb{E}}_{r \sim M(c)}\left[\log \frac{p_M(r|c)}{p(r)}\right]}_{\geq 0}. \tag{20}$$

The second term in Equation 20 is the **specific information** of the value c about the random variable $M(c)$. Specific information must be non-negative, as proven by Blachman (1968). Therefore we have:

$$\begin{aligned} \text{distortion}_M(w, c) &\leq \log \frac{1}{p(c)} \\ &= -\log p(c). \end{aligned}$$

□

Proposition 1 showed that listeners are more liable to make incorrect predictions based on lower-probability contexts. On average, these different predictions will manifest as increased difficulty, rather than decreased difficulty, as shown below in Proposition 2.

Now we turn to the second deduction: that noisiness in memory representations will make comprehenders regress to their prior expectations about words which they would have had regardless of context. Concretely, we show that the value of lossy-context surprisal is on average somewhere between full-context surprisal and unigram surprisal. We show that this is true in terms of the average predicted difficulties for words in contexts.

Proposition 2. *For all distributions L over contexts c and words w, and all distributions M over memory representations r given contexts, we have:*

$$\mathop{\mathbb{E}}_{c,w\sim L}\left[D_{\textit{surprisal}}(w|c)\right] \leq \mathop{\mathbb{E}}_{c,w\sim L}\left[D_{\textit{lc surprisal}}(w|c)\right] \leq \mathop{\mathbb{E}}_{w\sim L}\left[-\log p(w)\right].$$

Proof. We start by writing the expected surprisal values as entropy and conditional entropy values (Cover & Thomas, 2006):

$$\begin{aligned}
&\mathop{\mathbb{E}}_{c,w\sim L}\left[D_{\text{surprisal}}(w|c)\right] = \mathop{\mathbb{E}}_{c,w\sim L}\left[-\log p(w|c)\right] \equiv H[W|C] \\
&\mathop{\mathbb{E}}_{c,w\sim L}\left[D_{\text{lc surprisal}}(w|c)\right] = \mathop{\mathbb{E}}_{c,w\sim L,r\sim M(c)}\left[-\log p(w|r)\right] \equiv H[W|R] \\
&\mathop{\mathbb{E}}_{w\sim L}\left[-\log p(w)\right] \equiv H[W],
\end{aligned}$$

where $H[W|C]$ is the conditional entropy of words given contexts, $H[W|R]$ is the conditional entropy of word given memory representations, and $H[W]$ is the unigram entropy of words. Next we make use of a general information-theoretic result called the Shannon Inequality, which holds for all random variables X and Y:

$$H[X|Y] \leq H[X],$$

that is, the conditional entropy of X given Y is always less than or equal to the unconditional entropy of X. First we have:

$$H[W|R] \leq H[W],$$

which establishes the second inequality in the proposition. Next, we use the fact that words are conditionally independent from memory representations given contexts ($W \perp R|C$) to write:

$$H[W|C] = H[W|C,R].$$

Using the Shannon Inequality again, we have:

$$H[W|C] = H[W|C,R] \leq H[W|R],$$

which establishes the first inequality in the proposition. □